SLiMSearch 2.0: biological context for short linear motifs in proteins

Norman E. Davey¹, Niall J. Haslam², Denis C. Shields²* and Richard J. Edwards³

¹Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, ²UCD Complex and Adaptive Systems Laboratory, UCD Conway Institute, and School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Ireland and ³School of Biological Sciences, University of Southampton, Southampton, UK

Received February 12, 2011; Revised April 26, 2011; Accepted May 5, 2011

ABSTRACT

Short, linear motifs (SLiMs) play a critical role in many biological processes. The SLiMSearch 2.0 (Short, Linear Motif Search) web server allows researchers to identify occurrences of a user-defined SLiM in a proteome, using conservation and protein disorder context statistics to rank occurrences. User-friendly output and visualizations of motif context allow the user to quickly gain insight into the validity of a putatively functional motif occurrence. For each motif occurrence, overlapping UniProt features and annotated SLiMs are displayed. Visualization also includes annotated multiple sequence alignments surrounding each occurrence, showing conservation and protein disorder statistics in addition to known and predicted SLiMs, protein domains and known post-translational modifications. In addition, enrichment of Gene Ontology terms and protein interaction partners are provided as indicators of possible motif function. All web server results are available for download. Users can search motifs against the human proteome or a subset thereof defined by Uniprot accession numbers or GO term. The SLiMSearch server is available at: http://bioware.ucd.ie/slimsearch2.html.

INTRODUCTION

The purpose of the SLiMSearch (Short, Linear Motif Search) web server is to allow researchers to identify novel occurrences of user-defined Short Linear Motifs (SLiMs) in a set of sequences. SLiMs, also referred to as linear motifs or minimotifs, are functional microdomains that play a central role in many diverse biological pathways through post-translational modification (including cleavage), subcellular localization and ligand binding. Once a SLiM has been defined, finding matches in a given set of protein sequences is a fairly trivial task. Several web-based methods to discover novel instances of known SLiMs are available, including ELM (2), MnM (3), SIRW (4), ScanProsite (5) and QuasiMotifFinder (6), which generally utilize databases of known motif patterns to search query protein sequences supplied by the user.

While finding matches is trivial, however, interpreting their biological significance is far from easy. Stochastic occurrences of small, degenerate motifs are common; distinguishing real occurrences from the background of random motif hits remains the greatest challenge in a priori motif discovery. One approach is to simply filter out motifs that are likely to occur numerous times by chance—ScanProsite (5), for example, has an option to ‘Exclude motifs with a high probability of occurrence’, while QuasiMotifFinder (6) uses the background occurrence of motifs in PfamA families (7) to assess the significance of hits. These strategies work well for longer, family descriptor motifs, but are not so useful for SLiMs because of their tendency to occur by chance. Instead, additional contextual information such as sequence conservation (3,6,9,10), structural context (3,11) or even biological keywords (4) can be used to assess the likelihood of true functional significance for putatively functional sites.

Most motif search tools rely on pre-existing motif libraries, such as ELM (2), MnM (3) or Prosite (8). Those that permit users to define their own motifs, such as ScanProsite (5) and QuasiMotifFinder but are not so useful for SLiMs because of their tendency to occur by chance. Instead, additional contextual information such as sequence conservation (3,6,9,10), structural context (3,11) or even biological keywords (4) can be used to assess the likelihood of true functional significance for putatively functional sites.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
SLiMFinder (16) and FIRE-pro (17). Although SLiMFinder (16) estimates the statistical significance of returned motif predictions, correcting for biases introduced by evolutionary relationships within the data, assessing the ‘biological’ significance of predicted SLiMs remains challenging. One approach is to compare candidate SLiMs to existing motif libraries to identify similarities to previously known motifs (18). When a genuinely novel motif is predicted, however, knowledge of existing motifs is of limited use. Instead, it is useful to be able to establish the background distribution of occurrences of the novel motif, utilizing contextual information to help screen out the inevitable spurious chance matches.

We recently made our powerful de novo SLiM discovery tool, SLiMFinder (16), available as a web server (19). To aid interpretation of SLiMFinder results, we made a new tool available, SLiMSearch, which allows users to search protein data sets with user-defined motifs, including motif prediction output from SLiMFinder (20). SLiMSearch utilized the same sequence context assessment as SLiMFinder, enabling results to be masked or ranked based on the important biological indicators of sequence conservation and structural disorder (19, 21) and features the same SLiMChance algorithm for assessing statistical overrepresentation of SLiM occurrences, correcting for biases introduced by evolutionary relationships within the data (16). Like SLiMFinder, SLiMSearch was optimized for small protein data sets. In this article, we describe a complementary server, SLiMSearch 2.0, which is optimized for searches of a whole proteome.

SLiMSearch 2.0 replaces SLiMChance data set probabilities with individual likelihoods for each motif instance that permit the ranking of many motif occurrences and helps separate putative functional instances from the background of stochastic occurrences. A comprehensive study of the Eukaryotic Linear Motif (ELM) database by Fuxreiter et al. (22) found that SLiMs are more likely to be found in disordered regions, while Chica et al. (9) found that conserved motifs are more likely to be true positives. Our previous work with both discovery of new instances of known motifs and of novel motifs shows that, motifs in disordered regions and conserved motifs are typically (but not always) more likely to be true positives (10). Therefore, we encourage the use of an optional disorder filter and we present the results ranked according to conservation. Enrichment scores for motif counts are calculated (i) versus reversed/shuffled variants of the motif, (ii) for Gene Ontology (GO) terms (23) and (iii) for known BioGRID interactors of individual hub proteins (24). In addition to identifying individual occurrences of known motifs, therefore, SLiMSearch 2.0 can indicate possible functional significance for entirely novel motifs. Input, output and results visualizations are fully compatible with our existing SLiM analysis web servers, SLiMDisc (25), CompariMotif (18), SLiMFinder (19) and SLiMSearch 1.0 (20), providing a suite of integrated tools for analysing these biologically important sequence features.

**THE SLIMSEARCH 2.0 ALGORITHM**

SLiMSearch 2.0 performs a motif regular expression search against a proteome allowing restriction of considered sequences to set of proteins or a given GO term. Features include annotation of overlapping sequence annotation and calculation of global and local motif statistics and attributes.

**Pre-formatted database**

To speed up motif attribute calculations, pre-computed databases for each proteome are used. The current release has only Human UniProt release v1.37 (Aug 2010) (26); however, more model proteomes will be added as data is computed. Two pre-computed conservation scores are calculated for each protein in the proteome, a column-based tree-weighted conservation score (WCS) (9) and a relative local conservation (RLC) metric (10). Homologues for each sequence are identified using a BLAST search against a database of 70 complete Ensembl proteomes (Ensembl 59, October 2010, 69 Metazoan proteomes and *Saccharomyces cerevisiae*) (27) and orthologues are predicted using GOPHER (default options) (25). Predicted orthologues are aligned by MAFFT (28) and used to calculate conservation scoring metrics on a residue-by-residue basis. Disorder scores for each residue are calculated using IUPred (default options) (21).

Several features of interest are also preformatted for rapid querying: (i) Domain data from Pfam (29); (ii) structure data from PDB (30); (iii) experimentally validated motifs from the ELM database (2); and (iv) SNP and modification data from UniProt annotation (26).

**Scoring**

The IUPred disorder score, IUP, of the motif is calculated as the mean disorder score across the defined (non-wildcard) residues. The WCS of a motif is calculated similarly. SLiMSearch 2.0 extends the RLC score to return a probability. Based on the assumption, consistent with empirical observation, that the RLC scores for a residue are normally distributed (10), the RLC of a residue is converted into a probability, $P(\text{RLC})$, using the Gaussian Cumulative Distribution Function (CDF). The relative conservation probability of a motif, $P$, the probability of each residue of a motif having its given RLC or higher can be calculated as the product of the $P(\text{RLC})$ for each residue within the motif. A significance value, $P(\text{cons})$, representing the probability of a given motif having that $P$-value or lower by chance, can then be calculated for the motifs $P$-value using the CDF of the uniform product distribution [Equation (1)]. Thus, the $P(\text{cons})$ statistic provides a useful measure of how likely it is that this motif will have the observed degree of local conservation (or higher) by chance. Note that it does not provide any indication of the probability of the motif itself, which is best inferred from the enrichment values.

$$P(\text{cons}) = \frac{(-1)^n(-\ln(P))^{-n}\ln(P)^n\Gamma(n, -\ln(P))}{(n - 1)!}$$

(1)
Enrichment scores

Enrichment scores for motif counts are calculated for the input motif against the reverse of the motif and a randomly shuffled variant of the motif. The score is a simple quotient, where the input motif count is the divisor and the shuffled or reverse motif count is the dividend. Enrichment scores for each GO term and BioGRID interaction hub protein are calculated versus the expectation provided by the whole proteome, i.e. the number of motif occurrences in proteins with that GO term/interaction partner divided by the expected number of proteins, which is the total number of proteins with the motif multiplied by the proportion of the proteome with that GO term/interaction partner. Enrichment significance is calculated using the Fisher’s exact test. Counts are normalized for independence by clustering highly similar proteins based on UniRef50 groups.

THE SLIMSEARCH 2.0 WEBSERVER

The SLiMSearch 2.0 server is available at http://bioware.ucd.ie/slimsearch2.html. The website is free and open to all and there is no login requirement. The purpose of the web server is to allow researchers to identify novel occurrences of user-defined SLiMs in a set of sequences. A rapid pattern matching search is first performed to identify all occurrences of the motif in the proteome (or a defined subset). Pre-formatted databases are then used to rapidly extract scores and sequence features for each occurrence before enrichment scores are calculated. Interactive output and visualizations permit easy exploration of returned occurrences of the motif and their sequence context. These features of the web server are described in more detail in the following sections.

Input

A motif to be compared against the search is the sole compulsory input. The motif should be expressed as a regular expression using single letter amino acid codes (e.g. R.LF or RxLF but not Arg–x–Leu–Phe). The format is a simple quotient, where the input motif count is the divisor and the shuffled or reverse motif count is the dividend. Enrichment scores for each GO term and BioGRID interaction hub protein are calculated versus the expectation provided by the whole proteome, i.e. the number of motif occurrences in proteins with that GO term/interaction partner divided by the expected number of proteins, which is the total number of proteins with the motif multiplied by the proportion of the proteome with that GO term/interaction partner. Enrichment significance is calculated using the Fisher’s exact test. Counts are normalized for independence by clustering highly similar proteins based on UniRef50 groups.
Motif Hits

Top 10 by conservation (See alt.png)

<table>
<thead>
<tr>
<th>Accession</th>
<th>Gene</th>
<th>Name</th>
<th>p(con Ret)</th>
<th>WCS</th>
<th>UTP</th>
<th>Motif</th>
<th>RE</th>
<th>Start</th>
<th>Pfram</th>
<th>POB</th>
<th>Mod</th>
<th>SNP</th>
<th>ELM</th>
<th>Other</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>06P0J2</td>
<td>C14orf43</td>
<td>Uncharacterized protein C14orf43</td>
<td>0.003</td>
<td>0.8</td>
<td>0.52</td>
<td>HATOT</td>
<td>HR</td>
<td>TOT</td>
<td>969</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09Y5P3</td>
<td>RA2</td>
<td>Retinoic acid-induced protein 2</td>
<td>0.004</td>
<td>0.89</td>
<td>0.97</td>
<td>HGTOT</td>
<td>HR</td>
<td>TOT</td>
<td>277</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09S26T</td>
<td>RAGRPA1</td>
<td>RAS guanyl-releasing protein 1</td>
<td>0.006</td>
<td>0.87</td>
<td>0.47</td>
<td>HATOT</td>
<td>HR</td>
<td>TOT</td>
<td>671</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P42731</td>
<td>ARHSP25P2</td>
<td>Rho GTPase-activating protein 25</td>
<td>0.007</td>
<td>0.94</td>
<td>0.57</td>
<td>HRHTO</td>
<td>HR</td>
<td>TOT</td>
<td>438</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G13498</td>
<td>DYNCe1E</td>
<td>Cytoplasmic dynein 1 intermediate chain 2</td>
<td>0.01</td>
<td>0.95</td>
<td>0.59</td>
<td>HETOT</td>
<td>HR</td>
<td>TOT</td>
<td>158</td>
<td>Dynemin LC2 (192-164)</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04S313</td>
<td>ATMIN</td>
<td>ATM interactor</td>
<td>0.011</td>
<td>0.96</td>
<td>0.64</td>
<td>RETOT</td>
<td>HR</td>
<td>TOT</td>
<td>491</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G14576</td>
<td>DYNc1H1</td>
<td>Cytoplasmic dynein 1 intermediate chain 1</td>
<td>0.014</td>
<td>0.89</td>
<td>0.56</td>
<td>HETOT</td>
<td>HR</td>
<td>TOT</td>
<td>168</td>
<td>Dynemin LC2 (142-174)</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G07228</td>
<td>RAF5P3</td>
<td>RAF3-interacting JNK-activating modulator</td>
<td>0.026</td>
<td>0.83</td>
<td>0.82</td>
<td>RFTOT</td>
<td>HR</td>
<td>TOT</td>
<td>164</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G09161</td>
<td>ATT</td>
<td>Protein AATP</td>
<td>0.034</td>
<td>0.75</td>
<td>0.34</td>
<td>RHTOT</td>
<td>HR</td>
<td>TOT</td>
<td>408</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P46013</td>
<td>MKH3</td>
<td>Antigen [r]-67</td>
<td>0.043</td>
<td>0.54</td>
<td>0.52</td>
<td>LHTOT</td>
<td>HR</td>
<td>TOT</td>
<td>2017</td>
<td>K16TR (1976-2007)</td>
<td>view</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Top 5 enriched GO terms by enrichment significance (See alt.png)

| p(| enrichment | GO id | GO term | Proteins | p(| enrichment | GO id | GO term | Proteins |
|------------|-------|---------|----------|----------|------------|-------|---------|----------|
| 1.15e-06   | 113   | GO:0008218 | bioluminescence | 3 proteins | 1.000 | 5 | 10.348 | Dynemin light chain 1, cytoplasmic |
| 5.8e-05    | 3.57  | GO:0007165 | cell adhesion protein | 3 proteins | 1.000 | 3 | 28.384 | Dynemin light chain 2, cytoplasmic |
| 0.000179   | 26.5  | GO:0018298 | chromosome ligation | 3 proteins | 1.001 | 3 | 18.262 | E3 ubiquitin-protein ligase RFCD2 |
| 0.002388   | 8.83  | GO:0008544 | endosomal development | 5 proteins | 1.001 | 2 | 66.229 | Hsp70-binding protein 1 |
| 0.000895   | 42.9  | GO:0007130 | complex assembly | 2 proteins | 1.001 | 2 | 66.229 | Dynemin light chain Tctex-type 3 |

Motif Statistics

<table>
<thead>
<tr>
<th>Type</th>
<th>Motif</th>
<th>Enrichment Instances</th>
<th>Proteins</th>
<th>Dataset Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif</td>
<td>KR</td>
<td>TOT</td>
<td>-</td>
<td>161</td>
</tr>
<tr>
<td>Reverse</td>
<td>TQT</td>
<td>HR</td>
<td>0.957</td>
<td>154</td>
</tr>
<tr>
<td>shufle</td>
<td>HR</td>
<td>TOT</td>
<td>0.986</td>
<td>156</td>
</tr>
</tbody>
</table>

Top 5 enriched interactors by enrichment significance (See alt.png)

FUNDING

Science Foundation Ireland (08/IN.1/B1864) and the University of Southampton; European Molecular Biology Laboratory [EMBL Interdisciplinary Postdoc (EIPOD) fellowship to N.E.D.] Funding for open access charge: The University of Southampton.

Conflicts of interest statement. None declared.

REFERENCES


